

Sensitivity analysis, multilinearity and beyond

Manuele Leonelli

Departamento de Estatística, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brazil

Christiane Görgen and Jim Q. Smith

Department of Statistics, The University of Warwick, Coventry, UK

Abstract

Sensitivity methods for the analysis of the outputs of discrete Bayesian networks have been extensively studied and implemented in different software packages. These methods usually focus on the study of sensitivity functions and on the impact of a parameter change to the Chan-Darwiche distance. Although not fully recognized, the majority of these results rely heavily on the multilinear structure of atomic probabilities in terms of the conditional probability parameters associated with this type of network. By defining a statistical model through the polynomial expression of its associated defining conditional probabilities, we develop here a unifying approach to sensitivity methods applicable to a large suite of models including extensions of Bayesian networks, for instance context-specific and dynamic ones. Our algebraic approach enables us to prove that for models whose defining polynomial is multilinear both the Chan-Darwiche distance and any divergence in the family of ϕ -divergences are minimized for a certain class of multi-parameter contemporaneous variations when parameters are proportionally covaried.

Keywords: Bayesian networks, CD distance, Interpolating Polynomial, Sensitivity Analysis, ϕ -divergences.

1. Introduction

Many discrete statistical problems in a variety of domains are nowadays often modeled using *Bayesian networks* (BNs) [38]. There are now thousands of practical applications of these models [1, 7, 27, 29], which have spawned many useful technical

developments: including a variety of fast exact, approximate and symbolic propagation algorithms for the computation of probabilities that exploit the underlying graph structure [17, 19, 20]. Some of these advances have been hard-wired into software [9, 31, 34] which has further increased the applicability and success of these methods.

However, BN modeling would not have experienced such a widespread application without tailored methodologies of *model validation*, i.e. checking that a model produces outputs that are in line with current understanding, following a defensible and expected mechanism [22, 41]. Such techniques are now well established for BN models [15, 31, 41, 42]. These are especially fundamental for expert elicited models, where both the probabilities and the covariance structure are defined from the suggestions of domains experts, following knowledge engineering protocols tailored to the BN's building process [36, 43]. We can broadly break down the validation process into two steps: the first concerns the auditing of the underlying graphical structure; the second, assuming the graph represents a user's beliefs, checks the impact of the numerical elicited probabilities within this parametric family on outputs of interest. The focus of this paper lies in this second validation phase, usually called a *sensitivity analysis*.

The most common investigation is the so-called *one-way* sensitivity analysis, where the impacts of changes made to a single probability parameter are studied. Analyses where more than one parameter at a time is varied are usually referred to as *multi-way*. In both cases a complete sensitivity analysis for discrete BNs often involves the study of *Chan-Darwiche (CD) distances* [9, 10, 11] and *sensitivity functions* [16, 48]. The CD distance is used to quantify global changes. It measures how the overall distribution behaves when one (or more) parameter is varied. A significant proportion of research has focused on identifying parameter changes such that the original and the 'varied' BN distributions are close in CD distance [11, 45]. This is minimized when, after a single arbitrary parameter change, other covarying parameters, e.g. those from the same conditional distribution, have the same proportion of the residual probability mass as they originally had. Sensitivity functions, on the other hand, model local changes with respect to an output of interest. These describe how that output probability varies as one (or potentially more) parameter is allowed to be changed. Although both these concepts can be applied to generic Bayesian analyses, they have almost exclusively

been discussed and applied only within the BN literature (see [12, 13, 14, 44] for some exceptions). This is because the computations of both CD distances and sensitivity functions are particularly straightforward for BN models.

In this paper we introduce a unifying comprehensive framework for certain multi-way analyses, usually called in the context of BNs *single full conditional probability table (CPT) analyses* - where one parameter from each CPT of one vertex of a BN given each configurations of its parents is varied. Using the notion of an interpolating polynomial [40] we are able to describe a large variety of models based on their polynomial form. Then, given this algebraic characterization, we demonstrate that one-way sensitivity methods defined for BNs can be generalized to single full CPT analyses for any model whose interpolating polynomial is multilinear, for example context-specific BNs [6] and chain event graphs [47]. Because of both the lack of theoretical results justifying their use and the increase in computational complexity, multi-way methods have not been extensively discussed in the literature: see [5, 10, 24] for some exceptions. This paper aims at providing a comprehensive theoretical toolbox to start applying such analyses in practice.

Importantly, our polynomial approach enables us to prove that single full CPT analyses in any multilinear polynomial model are optimal under proportional covariation in the sense that the CD distance between the original and the varied distributions is minimized. The optimality of this covariation method has been an open problem in the sensitivity analysis literature for quite some time [10, 45]. However, we are able to provide further theoretical justifications for the use of proportional covariation in single full CPT analyses. We demonstrate below that for any multilinear model this scheme minimizes not only the CD distance, but also any divergence in the family of ϕ -divergences [2, 18]. The class of ϕ -divergences include a very large number of divergences and distances (see e.g. [37] for a review), including the famous Kullback-Leibler (KL) divergence [32]. The application of KL distances in sensitivity analyses of BNs has been almost exclusively restricted to the case when the underlying distribution is assumed Gaussian [23, 24], because in discrete BNs the computation of such a divergence requires more computational power than for CD distances. We will demonstrate below that this additional complexity is a feature shared by any divergence in the

family of ϕ -divergences.

However, by studying sensitivity analysis from a polynomial point of view, we are able to consider a much larger class of models for which such methods are very limited. We investigate the properties of one-way sensitivity analysis in models whose interpolating polynomial is not multilinear, which are usually associated to dynamic settings where probabilities are recursively defined. This difference gives us an even richer class of sensitivity functions as shown in [13, 14, 44] for certain dynamic BN models, which are not simply linear but more generally polynomial. We further introduce a procedure to compute the CD distance in these models and demonstrate that no unique updating of covarying parameters lead to the smallest CD distance between the original and the varied distribution.

The paper is structured as follows. In Section 2 we define interpolating polynomials and demonstrate that many commonly used models entertain a polynomial representation. In Section 3 we review a variety of divergence measures. Section 4 presents a variety of results for single full CPT sensitivity analyses in multilinear models. In Section 5 the focus moves to non-multilinear models and one-way analyses. We conclude with a discussion.

2. Multilinear and polynomial parametric models

In this section we first provide a generic definition of a parametric statistical model together with the notion of interpolating polynomial. We then categorize parametric models according to the form of their interpolating polynomial and show that many commonly used models fall within two classes.

2.1. Parametric models and interpolating polynomials

Let $\mathbf{Y} = (Y_1, \dots, Y_m)$ be a random vector with an associated discrete and finite sample space \mathbb{Y} , with $\#\mathbb{Y} = n$. Although our methods straightforwardly applies when the entries of \mathbf{Y} are random vectors, for ease of notation, we henceforth assume its elements are univariate.

Definition 1. Denote by $\mathbf{p}_\theta = (p_\theta(\mathbf{y}) \mid \mathbf{y} \in \mathbb{Y})$ the vector of values of a probability mass function $p_\theta : \mathbb{Y} \rightarrow [0, 1]$ which depends on a choice of parameters $\theta \in \mathbb{R}^k$. The entries of \mathbf{p}_θ are called *atomic probabilities* and the elements $\mathbf{y} \in \mathbb{Y}$ *atoms*.

Definition 2. A discrete *parametric statistical model* on $n \in \mathbb{N}$ atoms is a subset $\mathbb{P}_\Psi \subseteq \Delta_{n-1}$ of the $n - 1$ dimensional probability simplex, where

$$\Psi : \mathbb{R}^k \rightarrow \mathbb{P}_\Psi, \theta \mapsto \mathbf{p}_\theta, \quad (1)$$

is a bijective map identifying a particular choice of parameters $\theta \in \mathbb{R}^k$ with one vector of atomic probabilities. The map Ψ is called a *parametrisation* of the model.

The above definition is often encountered in the field of *algebraic statistics*, where properties of statistical models are studied using techniques from algebraic geometry and commutative computer algebra, among others [21, 46]. We next follow [25] in extending some standard terminology.

Definition 3. A model $\mathbb{P}_\Psi \subseteq \Delta_{n-1}$ has a *monomial parametrisation* if

$$p_\theta(\mathbf{y}) = \theta^{\alpha_{\mathbf{y}}}, \quad \text{for all } \mathbf{y} \in \mathbb{Y},$$

where $\alpha_{\mathbf{y}} \in \mathbb{N}_0^k$ denotes a vector of exponents and $\theta^{\alpha_{\mathbf{y}}} = \theta_1^{\alpha_{1,\mathbf{y}}} \cdots \theta_k^{\alpha_{k,\mathbf{y}}}$ is a monomial. Then equation (1) is a monomial map and $\theta^{\alpha_{\mathbf{y}}} \in \mathbb{R}_k[\Theta]$, for all $\mathbf{y} \in \mathbb{Y}$. Here $\Theta = \{\theta_1, \dots, \theta_k\}$ is the set of indeterminates and $\mathbb{R}_k[\Theta]$ is the polynomial ring over the field \mathbb{R} .

For models entertaining a monomial parametrisation the network polynomial we introduce in Definition 4 below concisely captures the model structure and provides a platform to answer inferential queries [20, 26].

Definition 4. The *network polynomial* of a model \mathbb{P}_Ψ with monomial parametrisation Ψ is given by

$$c_{\mathbb{P}_\Psi}(\theta, \lambda) = \sum_{\mathbf{y} \in \mathbb{Y}} \lambda_{\mathbf{y}} \theta^{\alpha_{\mathbf{y}}},$$

where $\lambda_{\mathbf{y}}$ is an indicator function for the atom \mathbf{y} .

Probabilities of events in the underlying sigma-field can be computed from the network polynomial by setting equal to one the indicator function of atoms associated to that event. In the following it will be convenient to work with a special case of the network polynomial where all the indicator functions are set to one.

Definition 5. The *interpolating polynomial* of a model \mathbb{P}_Ψ with monomial parametrisation Ψ is given by the sum of all atomic probabilities,

$$c_{\mathbb{P}_\Psi}(\theta) = \sum_{\alpha \in \mathbb{A}} \theta^\alpha,$$

where $\mathbb{A} = \{\alpha_y \mid y \in \mathbb{Y}\} \subset \mathbb{N}_0^k$.

2.2. Multilinear models

In this work we will mostly focus on parametric models whose interpolating polynomial is multilinear.

Definition 6. We say that a parametric model \mathbb{P}_Ψ is *multilinear* if its associated interpolating polynomial is multilinear, i.e. if $\mathbb{A} \subseteq \{0, 1\}^k$.

We note here that a great portion of well-known non-dynamic graphical models are multilinear. We explicitly show below that this is the case for BNs and context-specific BNs [6]. In [26] we showed that certain chain event graph models [47] have multilinear interpolating polynomial. In addition, decomposable undirected graphs and probabilistic chain graphs [33] can be defined to have a monomial parametrisation whose associated interpolating polynomial is multilinear. An example of models non entertaining a monomial parametrisation in terms of atomic probabilities are non-decomposable undirected graphs, since their joint distribution can be written as a rational function of multilinear functions [12].

2.2.1. Bayesian networks

For an $m \in \mathbb{N}$, let $[m] = \{1, \dots, m\}$. We denote with Y_i , $i \in [m]$, a generic discrete random variable and with $\mathbb{Y}_i = \{0, \dots, m_i\}$ its associated sample space. For an $A \subseteq [m]$, we let $\mathbf{Y}_A = (Y_i)_{i \in A}$ and $\mathbb{Y}_A = \times_{i \in A} \mathbb{Y}_i$. Recall that for three random vectors \mathbf{Y}_i , \mathbf{Y}_j and \mathbf{Y}_k , we say that \mathbf{Y}_i is conditional independent of \mathbf{Y}_j given \mathbf{Y}_k , and write $\mathbf{Y}_i \perp\!\!\!\perp \mathbf{Y}_j \mid \mathbf{Y}_k$, if $\Pr(\mathbf{Y}_i = \mathbf{y}_i \mid \mathbf{Y}_j = \mathbf{y}_j, \mathbf{Y}_k = \mathbf{Y}_k) = \Pr(\mathbf{Y}_i = \mathbf{y}_i \mid \mathbf{Y}_k = \mathbf{Y}_k)$.

Definition 7. A BN over a discrete random vector $\mathbf{Y}_{[m]}$ consists of

- $m - 1$ *conditional independence* statements of the form $Y_i \perp\!\!\!\perp \mathbf{Y}_{[i-1]} \mid \mathbf{Y}_{\Pi_i}$, where $\Pi_i \subseteq [i - 1]$;
- a *directed acyclic graph* (DAG) \mathcal{G} with vertex set $V(\mathcal{G}) = \{Y_i : i \in [m]\}$ and edge set $E(\mathcal{G}) = \{(Y_i, Y_j) : j \in [m], i \in \Pi_j\}$;
- conditional probabilities $\theta_{i_j\pi} = \Pr(Y_i = j \mid \mathbf{Y}_{\Pi_i} = \boldsymbol{\pi})$ for every $j \in \mathbb{Y}_i$, $\boldsymbol{\pi} \in \mathbb{Y}_{\Pi_i}$ and $i \in [m]$.

The vector \mathbf{Y}_{Π_i} , $i \in [m]$, includes the *parents* of the vertex Y_i , i.e. those vertices Y_j such that there is an edge (Y_j, Y_i) in the DAG \mathcal{G} of the BN.

From [10] we know that for any atom $\mathbf{y} \in \mathbb{Y}_{[m]}$ its associated monomial in the network polynomial can be written as

$$p_{\theta}(\mathbf{y}) = \prod_{\mathbf{y} \sim \{i_j, \boldsymbol{\pi}\}} \lambda_{i_j} \theta_{i_j\boldsymbol{\pi}},$$

where \sim denotes the compatibility relation among instantiations.

Lemma 1. *From [20, 26], the interpolating polynomial of a BN model can be written as*

$$c_{\text{BN}}(\theta) = \sum_{\mathbf{y} \in \mathbb{Y}_{[m]}} \prod_{\mathbf{y} \sim \{i_j, \boldsymbol{\pi}\}} \theta_{i_j\boldsymbol{\pi}}. \quad (2)$$

From Equation (2) we can immediately deduce the following.

Proposition 1. *A BN is a multilinear parametric model, whose interpolating polynomial is homogeneous with monomials of degree m .*

Example 1. Suppose a newborn is at risk of acquiring a disease and her parents are offered a screening test (Y_1) which can be either positive ($Y_1 = 1$) or negative ($Y_1 = 0$). Given that the newborn can either severely ($Y_2 = 2$) or mildly ($Y_2 = 1$) contract the disease or remain healthy ($Y_2 = 0$), her parents can then decide whether or not to give her a vaccine to prevent a relapse ($Y_3 = 1$ and $Y_3 = 0$, respectively). We assume that the parents' decision about the vaccine does not depend on the screening test if the

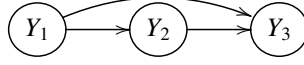


Figure 1: A BN model for the medical problem in Example 1.

newborn contracted the disease, and that the probability of being severely or mildly affected by the disease is equal for negative screening tests.

The above situation can be described, with some loss of information, by the BN in Figure 1, with probabilities, for $i, k \in \{0, 1\}$ and $j \in \{0, 1, 2\}$,

$$\Pr(Y_1 = i) = \theta_{1_i}, \quad \Pr(Y_2 = j | Y_1 = i) = \theta_{2_j 1_i}, \quad \Pr(Y_3 = k | Y_2 = j, Y_1 = i) = \theta_{3_k 2_j 1_i}.$$

Its associated interpolating polynomial has degree 3 and equals

$$c_{\text{BN}}(\theta) = \sum_{i=0}^1 \sum_{j=0}^2 \sum_{k=0}^1 \theta_{1_i} \theta_{2_j 1_i} \theta_{3_k 2_j 1_i}.$$

2.2.2. Context-specific Bayesian networks

In practice it has been recognized that often conditional independence statements do not hold over the whole sample space of certain conditioning variables but only for a subset of this, usually referred to as a *context*. A variety of methods have been introduced to embellish a BN with additional independence statements that hold only over contexts. A BN equipped with such embellishments is usually called *context-specific BN*. Here we consider the representation known as *context specific independence (CSI)-trees* and introduced in [6].

Example 2. Consider the medical problem in Example 1. Using the introduced notation, we notice that by assumption, for each $k = 0, 1$, the probabilities $\theta_{3_k 2_1 1_i}$ are equal for all $i = 0, 1$ and called $\theta_{3_k 2_1}$. Similarly, $\theta_{3_k 2_2 1_i}$ are equal and called $\theta_{3_k 2_2}$, $i, k = 0, 1$. Also $\theta_{2_2 1_0} = \theta_{2_1 1_0}$ are equal and called $\theta_{2_1 0}$. The first two constraints can be represented by the CSI-tree in Figure 2, where the inner nodes are random variables and the leaves are entries of the CPTs of one vertex. The tree shows that, if $Y_2 = 1$ or $Y_2 = 2$ then no matter what the value of Y_1 is, the CPT for $Y_3 = k$ will be equal to $\theta_{3_k 2_1}$ and $\theta_{3_k 2_2}$ respectively. The last constraint cannot be represented by a CSI-tree and is usually referred to as a *partial independence* [39]. In our polynomial approach, both partial and

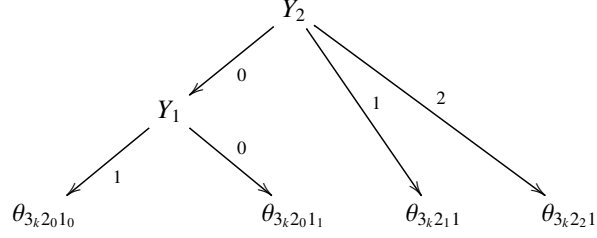


Figure 2: CSI-tree associated to vertex Y_3 of the BN in Figure 1 of Example 1, where $\theta_{3k2_11} = \Pr(Y_3 = k|Y_2 = 1)$, $\theta_{3k2_01_1} = \Pr(Y_3 = k|Y_2 = 0, Y_1 = 1)$ and $\theta_{3k2_01_0} = \Pr(Y_3 = k|Y_1 = 0, Y_2 = 0) = \Pr(Y_3|Y_1 = 2, Y_2 = 0)$.

context-specific independences can be straightforwardly imposed in the interpolating polynomial representation of the model. In fact the interpolating polynomial for the model in this example corresponds to the polynomial in equation (2) where the appropriate indeterminates are substituted with θ_{3k2_11} , θ_{3k2_21} and θ_{21_0} . This polynomial is again multilinear and homogeneous, just like for all context-specific BNs embellished with CSI-trees and partial independences.

We notice here that the interpolating polynomial of a multilinear model is not necessarily homogenous, as for example the one associated to certain chain event graph models, as shown in [26].

2.3. Non-multilinear models

Having discussed multilinear models, we now introduce more general structures which are often encountered in dynamic settings. Although many more models have this property, for instance dynamic chain graphs [3] and dynamic chain event graphs [4], for the purposes of this paper we focus here on the most commonly used model class of *dynamic Bayesian networks* (DBNs) [35]. In [26] we showed that the so-called non square-free chain event graph is also a non-multilinear model.

2.3.1. Dynamic Bayesian networks

DBNs extend the BN framework to dynamic and stochastic domains. As often in practice, we consider only stationary, feed-forward DBNs respecting the first order

Markov assumption with a finite horizon $T \in \mathbb{N}$, see e.g. [30]. This assumes that probabilities do not vary when shifted on time (stationarity), that current states only depend on the previous time point (first-order Markov assumption) and that contemporaneous variable cannot directly affect each other (feed-forward). These DBNs can be simply described by an initial distribution over the first time point and a BN having as vertex set two generic time slices. Such latter BN is usually called 2-Time slice Bayesian Network (2-TBN). Let $\{Y(t)\}_{t \in [T]} = \{Y_i(t) : i \in [m]\}_{t \in [T]}$ be a time series.

Definition 8. A 2-TBN for a time series $\{Y(t)\}_{t \in [T]}$ is a BN with DAG \mathcal{G} such that $V(\mathcal{G}) = \{Y_i(t), Y_i(t+1) : i \in [m]\}$ and its edge set is such that there are no edges $(Y_i(r), Y_j(r)), i, j \in [m], r = t, t+1, t \in [T-1]$.

Definition 9. A DBN for a time series $\{Y(t)\}_{t \in [T]}$ is a pair $(\mathcal{G}, \mathcal{G}')$, such that \mathcal{G} is a BN with vertex set $V(\mathcal{G}) = \{Y_i(1) : i \in [m]\}$, and \mathcal{G}' is a 2-TBN such that its vertex set $V(\mathcal{G}')$ is equal to $\{Y_i(t), Y_i(t+1) : i \in [m]\}$.

Example 3. Consider the problem of Example 1 and suppose the newborn can acquire the disease once a year. Suppose further that the screening test and the vaccine are available for kids up to four years old. This scenario can be modeled by a DBN with time horizon $T = 4$, where $Y_i(t), i \in [3], t \in [4]$, corresponds to the variable Y_i of Example 1 measured in the t -th year. Suppose that the probabilities of parents choosing the screening test and vaccination depend on whether or not the newborn acquired the disease in the previous year only. Furthermore, there is evidence that kids have a higher chance of contracting the disease if they were sick the previous year, whilst a lower chance if vaccination was chosen. This situation can be described by the DBN in Figure 3 where at time $t = 1$ the correlation structure of the non-dynamic problem is assumed.

For a finite time horizon $T = 4$, the interpolating polynomial has $2^8 \cdot 3^4$ monomials each of degree 12. To show that this polynomial is not multilinear, consider the event that the screening test is always positive, that the parents always decline vaccination and that the newborn gets mildly sick in her first three years of life, denoted as $\mathbf{y}_T = (Y_1(t) = 1, Y_2(s) = 1, Y_3(t) = 0, t \in [4], s \in [3])$. Let the parameters for the first time

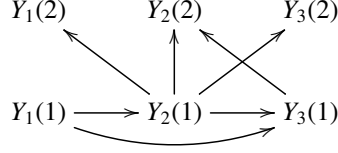


Figure 3: A DBN having at time $t = 1$ the DAG in Figure 1 and a 2-TBN with an edge from $Y_2(t)$ to $Y_i(t + 1)$, for $i \in [3]$, and the edge $(Y_3(t - 1), Y_2(t))$.

slice be denoted as in Example 1 and denote for $t = 2, 3, 4$

$$\begin{aligned}\hat{\theta}_{1,2_j} &= \Pr(Y_1(t) = i \mid Y_2(t - 1) = j), \quad i = 0, 1, \quad j = 0, 1, 2, \\ \hat{\theta}_{2,2_j,3_k} &= \Pr(Y_2(t) = i \mid Y_2(t - 1) = j, Y_3(t - 1) = k), \quad k = 0, 1, \quad i, j = 0, 1, 2, \\ \hat{\theta}_{3,2_j} &= \Pr(Y_3(t) = i \mid Y_2(t - 1) = j), \quad i = 0, 1, \quad j = 0, 1, 2.\end{aligned}$$

The interpolating polynomial for this event equals

$$c_{\text{DBN}}(\theta, \mathbf{y}_T) = \theta_{1_1} \theta_{2_1 1_1} \theta_{3_0 2_1 1_1} \hat{\theta}_{1_1 2_1}^3 \hat{\theta}_{3_0 2_1}^3 \left(\hat{\theta}_{2_1 2_1 3_0}^3 + \hat{\theta}_{2_1 2_1 3_0}^2 \hat{\theta}_{2_2 2_1 3_0} + \hat{\theta}_{2_1 2_1 3_0}^2 \hat{\theta}_{2_0 2_1 3_0} \right), \quad (3)$$

which has indeterminates of degree 3 and 2 and therefore is not multilinear.

Note that in the example above indeterminates can have degree up to $T - 1$, since this corresponds to the longest length of a path where the visited vertices can have probabilities that are identified in the ‘unrolled’ version of the DBN, i.e. one where the 2-TBN graph for time t is recursively collated to the one of time $t - 1$. From this observation the following follows.

Proposition 2. *A DBN is a parametric model with monomial parametrisation, whose interpolating polynomial is homogeneous and each indeterminate can have degree lower or equal to $T - 1$.*

As for multilinear models, the interpolating polynomial of a non-multilinear model can be non-homogeneous. This is the case for example for certain non square-free chain event graphs.

3. Divergence measures

In sensitivity analyses for discrete parametric statistical models we are often interested in studying how far apart from each other are two vectors of values of two probability mass functions \mathbf{p}_θ and $\mathbf{p}_{\bar{\theta}}$ from the same model \mathbb{P}_Ψ . Divergence measures are used to quantify this dissimilarity between probability distributions. In this section we provide a brief introduction to these functions within the context of our discrete parametric probability models.

Definition 10. A *divergence measure* \mathcal{D} within a discrete parametric probability model \mathbb{P}_Ψ is a function $\mathcal{D}(\cdot, \cdot) : \mathbb{P}_\Psi \times \mathbb{P}_\Psi \rightarrow \mathbb{R}$ such that for all $\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}} \in \mathbb{P}_\Psi$:

- $\mathcal{D}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) \geq 0$;
- $\mathcal{D}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) = 0$ iff $\mathbf{p}_\theta = \mathbf{p}_{\bar{\theta}}$.

The larger the divergence between two probability mass functions \mathbf{p}_θ and $\mathbf{p}_{\bar{\theta}}$, the more dissimilar these are. Notice that divergences are not formally metrics, since these do not have to be symmetric and respect the triangular inequality. We will refer to divergences with these two additional properties as *distances*.

The divergence most commonly used in practice is the KL divergence [32].

Definition 11. The *KL divergence* between $\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}} \in \mathbb{P}_\Psi$, $\mathcal{D}_{\text{KL}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}})$, is defined as

$$\mathcal{D}_{\text{KL}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) = \sum_{\mathbf{y} \in \mathbb{Y}} p_\theta(\mathbf{y}) \log \left(\frac{p_\theta(\mathbf{y})}{p_{\bar{\theta}}(\mathbf{y})} \right), \quad (4)$$

assuming $p_\theta(\mathbf{y}), p_{\bar{\theta}}(\mathbf{y}) > 0$ for all $\mathbf{y} \in \mathbb{Y}$.

Notice that the KL divergence is not symmetric and thus $\mathcal{D}_{\text{KL}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) \neq \mathcal{D}_{\text{KL}}(\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta)$ in general. However both divergences can be shown to be a particular instance of a very general family of divergences, called ϕ -divergences [2, 18].

Definition 12. The ϕ -divergence between $\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta \in \mathbb{P}_\Psi$, $\mathcal{D}_\phi(\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta)$, is defined as

$$\mathcal{D}_\phi(\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta) = \sum_{\mathbf{y} \in \mathbb{Y}} p_\theta(\mathbf{y}) \phi \left(\frac{p_{\bar{\theta}}(\mathbf{y})}{p_\theta(\mathbf{y})} \right), \quad \phi \in \Phi, \quad (5)$$

where Φ is the class of convex functions $\phi(x)$, $x \geq 0$, such that $\phi(1) = 0$, $0\phi(0/0) = 0$ and $0\phi(x/0) = \lim_{x \rightarrow \infty} \phi(x)/x$.

So for example $\mathcal{D}_{\text{KL}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) = \mathcal{D}_\phi(\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta)$ for $\phi(x) = -\log(x)$ and $\mathcal{D}_{\text{KL}}(\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta) = \mathcal{D}_\phi(\mathbf{p}_{\bar{\theta}}, \mathbf{p}_\theta)$ for $\phi(x) = x \log(x)$. Many other renowned divergences are in the family of ϕ -divergences: for example J divergences [28] and total variation distances (see [37] for a review).

The distance usually considered to study the dissimilarity of two probability mass functions in sensitivity analyses for discrete BNs is the aforementioned Chan-Darwiche distance. This distance is not a member of the ϕ -divergence family.

Definition 13. The *CD distance* between $\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}} \in \mathbb{P}_\Psi$, $\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}})$, is defined as

$$\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) = \log \max_{\mathbf{y} \in \mathbb{Y}} \frac{\mathbf{p}_{\bar{\theta}}(\mathbf{y})}{\mathbf{p}_\theta(\mathbf{y})} - \log \min_{\mathbf{y} \in \mathbb{Y}} \frac{\mathbf{p}_{\bar{\theta}}(\mathbf{y})}{\mathbf{p}_\theta(\mathbf{y})}, \quad (6)$$

where $0/0$ is defined as 1.

It has been noted that in sensitivity analysis in BNs, if one parameter of one CPT is varied, then the CD distance between the original and the varied BN equals the CD distance between the original and the varied CPT [11]. This distributive property, and its associated computational simplicity, has lead to a wide use of the CD distance in sensitivity studies in discrete BNs.

4. Sensitivity analysis in multilinear models

We can now formalize sensitivity analysis techniques for multilinear parametric models. We focus on an extension of single full CPT analyses from BNs to generic multilinear models. Standard one-way sensitivity analyses can be seen as a special case of single full CPT analyses when only one parameter is allowed to be varied. We demonstrate in this section that all the results about one-way sensitivity analysis in BN models extend to single full CPT analyses in multilinear parametric models and therefore hold under much weaker assumptions about the structure of both the sample space and the underlying conditional independences. Before presenting these results we review the theory of *covariation*.

4.1. Covariation

In one-way analyses one parameter within a parametrisation of a model is varied. When this is done, then *some* of the remaining parameters need to be varied as well to

respect the sum-to-one condition, so that the resulting measure is a probability measure. In the binary case this is straightforward, since the second parameter will be equal to one minus the other. But in generic discrete finite cases there are various considerations the user needs to take into account, as reviewed below.

Let $\theta_i \in \Theta$ be the parameter varied to $\tilde{\theta}_i$ and suppose this is associated to a random variable Y_C in the random vector \mathbf{Y} . Let $\Theta_C = \{\theta_1, \dots, \theta_i, \dots, \theta_r\} \subseteq \Theta$ be the subset of the parameter set including θ_i describing the probability distribution of Y_C and whose elements need to respect the sum to one condition. For instance Θ_C would include the entries of a CPT for a fixed combination of the parent variables in a BN model or the entries of a CPT associated to the conditional random variable from a leaf of a CSI-tree as in Figure 2. Suppose further these parameters are indexed according to their values, i.e. $\theta_1 \leq \dots \leq \theta_i \leq \dots \leq \theta_r$. From [45] we then have the following definition.

Definition 14. Let $\theta_i \in \Theta_C$ be varied to $\tilde{\theta}_i$. A *covariation* scheme $\sigma(\theta_j, \tilde{\theta}_i) : [0, 1]^2 \rightarrow [0, 1]$ is a function that takes as input the value of both $\tilde{\theta}_i$ and $\theta_j \in \Theta_C$ and returns an updated value for θ_j denoted as $\tilde{\theta}_j$.

Different covariation schemes may entertain different properties which, depending on the domain of application, might be more or less desirable. We now list some of these properties from [45].

Definition 15. In the notation of Definition 14, a covariation scheme $\sigma(\theta_j, \tilde{\theta}_i)$ is

- *valid*, if $\sum_{j \in [r]} \sigma(\theta_j, \tilde{\theta}_i) = 1$;
- *impossibility preserving*, if for any parameter $\theta_j = 0$, $j \neq i$, we have that $\sigma(\theta_j, \tilde{\theta}_i) = 0$;
- *order preserving*, if $\sigma(\theta_1, \tilde{\theta}_i) \leq \dots \leq \sigma(\theta_j, \tilde{\theta}_i) \leq \dots \leq \sigma(\theta_r, \tilde{\theta}_i)$;
- *identity preserving*, if $\sigma(\theta_j, \theta_i) = \theta_j$, $\forall j \in [r]$;
- *linear*, if $\sigma(\theta_j, \tilde{\theta}_i) = \gamma_j \tilde{\theta}_i + \delta_j$, for $\gamma_j \in [0, 1]$ and $\delta_j \in (-1, 1)$.

Of course any covariation scheme needs to be valid, otherwise the resulting measure is not a probability measure and any inference from the model would be misleading.

Applying a linear scheme is very natural: if for instance $\delta_j = -\gamma_j$, then $\sigma(\theta_j, \tilde{\theta}_i) = \delta_j(1 - \tilde{\theta}_i)$ and the scheme assigns a proportion δ_j of the remaining probability mass $1 - \tilde{\theta}_i$ to the remaining parameters. Following [45] we now introduce a number of frequently applied covariation schemes.

Definition 16. In the notation of Definition 14, we define

- the *proportional* covariation scheme, $\sigma_{\text{pro}}(\theta_j, \tilde{\theta}_i)$, as

$$\sigma_{\text{pro}}(\theta_j, \tilde{\theta}_i) = \begin{cases} \tilde{\theta}_i, & \text{if } j = i, \\ \frac{1 - \tilde{\theta}_i}{1 - \theta_i} \theta_j, & \text{otherwise.} \end{cases}$$

- the *uniform* covariation scheme, $\sigma_{\text{uni}}(\theta_j, \tilde{\theta}_i)$, for $r = \#\Theta_C$, as

$$\sigma_{\text{uni}}(\theta_j, \tilde{\theta}_i) = \begin{cases} \tilde{\theta}_i, & \text{if } j = i, \\ \frac{1 - \tilde{\theta}_i}{r - 1}, & \text{otherwise.} \end{cases}$$

- the *order preserving* covariation scheme, $\sigma_{\text{ord}}(\theta_j, \tilde{\theta}_i)$, for $i \neq r$, as

$$\sigma_{\text{ord}}(\theta_j, \tilde{\theta}_i) = \begin{cases} \tilde{\theta}_i, & \text{if } j = i, \\ \frac{\theta_j}{\theta_i} \tilde{\theta}_i, & \text{if } j < i \text{ and } \tilde{\theta}_i \leq \theta_i, \\ \frac{-\theta_j(1 - \theta_{\text{suc}})}{\theta_{\text{suc}}\theta_i} \tilde{\theta}_i + \frac{\theta_j}{\theta_{\text{suc}}}, & \text{if } j > i \text{ and } \tilde{\theta}_i \leq \theta_i, \\ \frac{\theta_j}{\theta_{\text{max}} - \theta_i} (\theta_{\text{max}} - \tilde{\theta}_i), & \text{if } j < i \text{ and } \tilde{\theta}_i > \theta_i, \\ \frac{\theta_j - \theta_{\text{max}}}{\theta_{\text{max}} - \theta_i} (\theta_{\text{max}} - \tilde{\theta}_i) + \theta_{\text{max}}, & \text{if } j > i \text{ and } \tilde{\theta}_i > \theta_i, \end{cases}$$

where $\theta_{\text{max}} = 1/(1 + r - i)$ is the upper bound for $\tilde{\theta}_i$ and $\theta_{\text{suc}} = \sum_{k=i+1}^r \theta_k$ is the original mass of the parameters succeeding θ_i in the ordering.

Table 1 summarizes which of the properties introduced in Definition 15 the above schemes entertain (see [45] for more details). Under proportional covariation, to all the covarying parameters is assigned the same proportion of the remaining probability mass as these originally had. Although this scheme is not order preserving, it maintains the order among the covarying parameters. The uniform scheme on the other hand gives the same amount of the remaining mass to all covarying parameters. In addition, although the order preserving scheme is the only one that entertains the order preserving property, this limits the possible variations allowed. Note that this scheme is

Scheme/Property	valid	imp-pres	ord-pres	ident-pres	linear
Proportional	✓	✓	✗	✓	✓
Uniform	✓	✗	✗	✗	✓
Order Preserving	✓	✓	✓	✓	✓

Table 1: Summary of the covariation schemes and the properties these entertain.

not only simply linear, but more precisely piece-wise linear, i.e. a function composed of straight-line sections. All the schemes in Definition 16 are domain independent and therefore can be applied with no prior knowledge about the application of interest. Other schemes, for instance domain dependent or non-linear, have been defined, but these are not of interest for the theory we develop here.

4.2. Sensitivity functions

We now generalize one-way sensitivity methods in BNs to the single full CPT case for general multilinear models. This type of analysis is simpler than other multi-way methods since the parameters varied/covared never appear in the same monomial of the BN interpolating polynomial. So we now find an analogous CPT analysis in multilinear models which has the same property. Suppose we vary n parameters $\theta_{1_i}, \dots, \theta_{n_i}$ and denote by $\Theta_j = \{\theta_{j_1}, \dots, \theta_{j_{r_j}}\}$, $j \in [n]$, the set of parameters including θ_{j_i} and associated to the same (conditional) random variable: thus respecting the sum to one condition. Assume these sets are such that $\cap_{j \in [n]} \Theta_j = \emptyset$. Note that a collection of such sets can not only be associated to the CPTs of one vertex given different parent configurations, but also, for instance, to the leaves of a CSI-tree as in Figure 2 or to the positions along the same *cut* in a CEG [47].

We start by investigating sensitivity functions. These describe the effect of the variation of the parameters $\theta_{1_i}, \dots, \theta_{n_i}$ on the probability of an event $\mathbb{Y}_T \subseteq \mathbb{Y}$ of interest. A sensitivity function $f_{\mathbb{Y}_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ equals the probability $\Pr(Y \in \mathbb{Y}_T) \triangleq p_{\tilde{\theta}}(\mathbf{y}_T)$ and is a function in $\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}$, where $\theta_{1_i}, \dots, \theta_{n_i}$ are varied to $\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}$. Our parametric definition of a statistical model enables us to explicitly express these as functions of the covariation scheme for any multilinear model. Recall that $\mathbb{A} = \{\alpha_{\mathbf{y}} | \mathbf{y} \in \mathbb{Y}\}$ and

Let $\mathbb{T} = \{\alpha_y \mid y \in \mathbb{Y}_T\}$. Let $\mathbb{A}_j, \mathbb{T}_j \subset \{0, 1\}^k$ be the subsets of \mathbb{A} and \mathbb{T} respectively including the exponents where the entry associated to an indeterminate in Θ_j is not zero, $\mathbb{A}_{j_s} \subseteq \mathbb{A}_j$ and $\mathbb{T}_{j_s} \subseteq \mathbb{T}_j$ be the subsets including the exponents such that the entry relative to θ_{j_s} is not zero, $j \in [n], s \in [r_j]$. Formally

$$\begin{aligned}\mathbb{A}_j &= \{\alpha_y \mid y \in \mathbb{Y}, \alpha_{j_s, y} \neq 0, s \in [r_j]\}, & \mathbb{T}_j &= \{\alpha_y \mid y \in \mathbb{Y}_T, \alpha_{j_s, y} \neq 0, s \in [r_j]\}, \\ \mathbb{A}_{j_s} &= \{\alpha_y \mid y \in \mathbb{Y}, \alpha_{j_s, y} \neq 0\}, & \mathbb{T}_{j_s} &= \{\alpha_y \mid y \in \mathbb{Y}_T, \alpha_{j_s, y} \neq 0\}.\end{aligned}$$

Let $\mathbb{A}_{-j_s}, \mathbb{T}_{-j_s} \subseteq \{0, 1\}^{k-1}$ be the sets including the elements in \mathbb{A}_{j_s} and \mathbb{T}_{j_s} , respectively, where the entry relative to $\theta_{j_s} \in \Theta_j$ is deleted. Lastly, let $\theta_{-j_s} = \prod_{\theta_k \in \Theta \setminus \{\theta_{j_s}\}} \theta_k$.

Proposition 3. *Consider a multilinear model \mathbb{P}_Ψ where the parameters $\theta_{j_i} \in \Theta_j$ are varied to $\tilde{\theta}_{j_i}$ and $\theta_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ is covaried according to a valid scheme $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$, $j \in [n], s \in [r_j] \setminus \{j_i\}$. The sensitivity function $f_{y_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ can then be written as*

$$f_{y_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}) = \sum_{j \in [n]} \sum_{\alpha \in \mathbb{T}_{-j_i}} \theta_{-j_i}^\alpha \tilde{\theta}_{j_i} + \sum_{j \in [n]} \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) + \sum_{\alpha \in \mathbb{T} \setminus \bigcup_{k \in [n]} \mathbb{T}_k} \theta^\alpha. \quad (7)$$

Proof. The probability of interest can be written as

$$\begin{aligned}p_\theta(y_T) &= \sum_{\alpha \in \mathbb{T}} \theta^\alpha = \sum_{j \in [n]} \sum_{s \in [r_j]} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \theta_{j_s} + \sum_{\alpha \in \mathbb{T} \setminus \bigcup_{k \in [n]} \mathbb{T}_k} \theta^\alpha \\ &= \sum_{j \in [n]} \sum_{\alpha \in \mathbb{T}_{-j_i}} \theta_{-j_i}^\alpha \theta_{j_i} + \sum_{j \in [n]} \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \theta_{j_s} + \sum_{\alpha \in \mathbb{T} \setminus \bigcup_{k \in [n]} \mathbb{T}_k} \theta^\alpha.\end{aligned}$$

The result follows by substituting the varying parameters with their varied version. \square

From Proposition 3 we can deduce that for a multilinear model, under a linear covariation scheme, the sensitivity function is multilinear.

Corollary 1. *Under the conditions of Proposition 3 and the linear covariation schemes $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) = \gamma_{j_s} \tilde{\theta}_{j_i} + \delta_{j_s}$, the sensitivity function $f_{y_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ equals*

$$f_{y_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}) = \sum_{j \in [n]} a_j \tilde{\theta}_{j_i} + b, \quad (8)$$

where

$$a_j = \sum_{\alpha \in \mathbb{T}_{-j_i}} \theta_{-j_i}^\alpha + \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \gamma_{j_s}, \quad b = \sum_{j \in [n]} \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \delta_{j_s} + \sum_{\alpha \in \mathbb{T} \setminus \bigcup_{k \in [n]} \mathbb{T}_k} \theta^\alpha. \quad (9)$$

Proof. The result follows by substituting the definition of a linear covariation scheme into equation (7) and then rearranging. \square

Therefore, under a linear covariation scheme, the sensitivity function is a multilinear function of the varying parameters $\tilde{\theta}_{j_i}$, $j \in [n]$. This was long known for BN models [8, 45, 48]. However, we have proven here that this feature is shared amongst all models having a multilinear interpolating polynomial. In BNs the computation of the coefficients a_j and b is particularly fast since for these models computationally efficient propagation techniques have been established. But these exist, albeit sometimes less efficiently, for other models as well (see e.g. [17] for chain graphs). Within our symbolic definition, we note however that once the exponent sets \mathbb{T}_{-j_s} , $s \in [r_j]$, are identified, then one can simply plug-in the values of the indeterminates to compute these coefficients.

We now deduce the sensitivity function when parameters are varied using the popular proportional scheme.

Corollary 2. *Under the conditions of Proposition 3 and proportional covariation scheme $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) = \frac{1-\tilde{\theta}_{j_i}}{1-\theta_{j_i}}\theta_{j_s}$, the sensitivity function, $f_{y_T}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ can be written in the multilinear form of equation (8), where*

$$a_j = \sum_{\alpha \in \mathbb{T}_{-j_i}} \theta_{-j_i}^\alpha - \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{j_s}} \frac{\theta^\alpha}{1 - \theta_{j_i}}, \quad b = \sum_{j \in [n]} \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{j_s}} \frac{\theta^\alpha}{1 - \theta_{j_i}} + \sum_{\alpha \in \mathbb{T} \setminus \bigcup_{k \in [n]} \mathbb{T}_k} \theta^\alpha.$$

Proof. For a proportional scheme the coefficients in the definition of a linear scheme equals $\gamma_{j_s} = -\theta_{j_s}/(1 - \theta_{j_i})$ and $\delta_{j_s} = \theta_{j_s}/(1 - \theta_{j_i})$. By substituting these expressions into equation (9) we have that

$$a_j = \sum_{\alpha \in \mathbb{T}_{-j_i}} \theta_{-j_i}^\alpha - \sum_{s \in [r_j] \setminus \{j_i\}} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \frac{\theta_{j_s}}{1 - \theta_{j_i}}, \quad b = \sum_{\substack{j \in [n] \\ s \in [r_j] \setminus \{j_i\}}} \sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \frac{\theta_{j_s}}{1 - \theta_{j_i}} + \sum_{\alpha \in \mathbb{T} \setminus \bigcup_{k \in [n]} \mathbb{T}_k} \theta^\alpha.$$

By noting that $\sum_{\alpha \in \mathbb{T}_{-j_s}} \theta_{-j_s}^\alpha \theta_{j_s} = \sum_{\alpha \in \mathbb{T}_{j_s}} \theta^\alpha$ the result then follows. \square

It is often of interest to investigate the posterior probability of a target event ($Y \in \mathbb{Y}_T$) given that an event ($Y \in \mathbb{Y}_O$) has been observed, $\mathbb{Y}_T, \mathbb{Y}_O \subseteq \mathbb{Y}$. This can be represented by the *posterior* sensitivity function $f_{y_T}^{y_O}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ describing the probability $\Pr(Y \in \mathbb{Y}_T \mid Y \in \mathbb{Y}_O)$ as a function of the varying parameters $\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}$.

$\theta_{1_1} = 0.4,$	$\theta_{2_2 1_1} = 0.3,$	$\theta_{2_1 1_1} = 0.5,$	$\theta_{2_1 0} = 0.3,$
$\theta_{3_0 2_2 1} = 0.1,$	$\theta_{3_0 2_1 1} = 0.3,$	$\theta_{3_0 2_0 1_1} = 0.7,$	$\theta_{3_0 2_0 1_0} = 0.8.$

Table 2: Probability specifications for Example 2.

Corollary 3. *Under the conditions of Corollary 1, a posterior sensitivity function $f_{y_T}^{y_O}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i})$ can be written as the ratio*

$$f_{y_T}^{y_O}(\tilde{\theta}_{1_i}, \dots, \tilde{\theta}_{n_i}) = \frac{\sum_{j \in [n]} a_j \tilde{\theta}_{j_i} + b}{\sum_{j \in [n]} c_j \tilde{\theta}_{j_i} + d}, \quad (10)$$

where $a_j, c_j \in [0, 1]$, $j \in [n]$, and $b, d \in (-1, 1)$.

Proof. The result follows from equation (8) and by noting that $\Pr(\mathbf{Y} \in \mathbb{Y}_T | \mathbf{Y} \in \mathbb{Y}_O) = \Pr(\mathbf{Y} \in \{\mathbb{Y}_T \cap \mathbb{Y}_O\}) / \Pr(\mathbf{Y} \in \mathbb{Y}_O)$. \square

The form of the coefficients in Corollary 3 can be deduced by simply adapting the notation of equation (7) to the events $\Pr(\mathbf{Y} \in \{\mathbb{Y}_T \cap \mathbb{Y}_O\})$ and $\Pr(\mathbf{Y} \in \mathbb{Y}_O)$ for the numerator and the denominator, respectively, of equation (10). Sensitivity functions describing posterior probabilities in BNs have been proven to entertain the form in equation (10). Again, Corollary 3 shows that this is so for any model having a multi-linear interpolating polynomial.

Example 4. Suppose the context-specific model definition in Example 2 is completed by the probability specifications in Table 2. Suppose we are interested in the event that parents do not decide for vaccination. Figure 4 shows the sensitivity functions for this event when $\theta_{2_1 1_1}$ (on the x-axis) and $\theta_{2_1 0}$ (on the y-axis) are varied and the other covarying parameter are changed with different schemes. We can notice that for all schemes the functions are linear in their arguments and that more precisely for an order-preserving scheme the sensitivity function is piece-wise linear. Notice that whilst uniform and proportional covariation assigns similar, although different, probabilities to the event of interest, under order-preserving covariation the probability of interest changes very differently from the other schemes: a property we have often observed in our investigations.

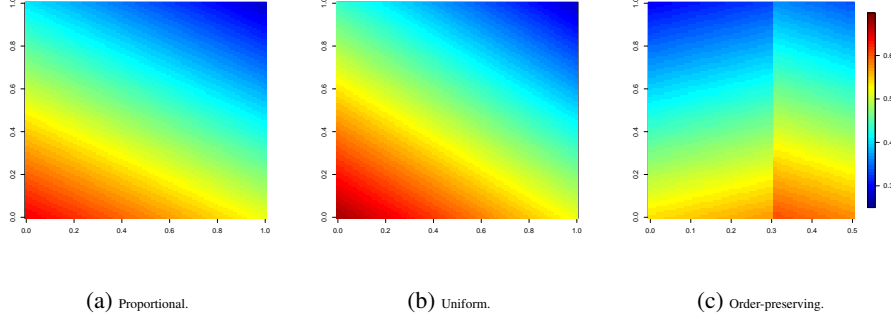


Figure 4: Sensitivity functions for Example 4 under different covariation schemes.

4.3. The Chan-Darwiche distance

Whilst sensitivity functions study local changes, CD distances describe global variations in distributions [11]. These can be used to study by how much two vectors of atomic probabilities vary in their distributional assumptions if one arises from the other via a covariation scheme. We are then interested in the global impact of that local change.

We next characterize the form of the CD distance for multilinear models in single full CPT analyses, first generalizing its form, again derived in [45] for BN models. We demonstrate that the distance depends only on the varied and covaried parameters: thus very easy to compute.

Proposition 4. *Let $\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}} \in \mathbb{P}_\Psi$, where \mathbb{P}_Ψ is a multilinear parametric model and $\mathbf{p}_{\tilde{\theta}}$ arises from \mathbf{p}_θ by varying θ_{j_i} to $\tilde{\theta}_{j_i}$ and $\theta_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ to $\tilde{\theta}_{j_s} = \sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$, where $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$ is a valid covariation scheme, $j \in [n]$, $s \in [r_j] \setminus \{j_i\}$. Then the CD distance between \mathbf{p}_θ and $\mathbf{p}_{\tilde{\theta}}$ is equal to*

$$\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log \max_{\substack{j \in [n] \\ s \in [r_j]}} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}} - \log \min_{\substack{j \in [n] \\ s \in [r_j]}} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}. \quad (11)$$

Proof. For a multilinear parametric model the CD distance can be written as

$$\begin{aligned}\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) &= \log \max_{\alpha \in \mathbb{A}} \frac{\tilde{\theta}^\alpha}{\theta^\alpha} - \log \min_{\alpha \in \mathbb{A}} \frac{\tilde{\theta}^\alpha}{\theta^\alpha} \\ &= \log \max \left\{ \max_{\substack{j \in [n] \\ \alpha \in \mathbb{A}_j}} \frac{\tilde{\theta}^\alpha}{\theta^\alpha}, \max_{\alpha \in \mathbb{A} \setminus \bigcup_{k \in [n]} \mathbb{A}_k} \frac{\tilde{\theta}^\alpha}{\theta^\alpha} \right\} - \log \min \left\{ \min_{\substack{j \in [n] \\ \alpha \in \mathbb{A}_j}} \frac{\tilde{\theta}^\alpha}{\theta^\alpha}, \min_{\alpha \in \mathbb{A} \setminus \bigcup_{k \in [n]} \mathbb{A}_k} \frac{\tilde{\theta}^\alpha}{\theta^\alpha} \right\}.\end{aligned}$$

If $\alpha \in \mathbb{A} \setminus \bigcup_{k \in [n]} \mathbb{A}_k$, then $\tilde{\theta} = \theta$ and thus $\tilde{\theta}^\alpha / \theta^\alpha = 1$. Because of the validity of the covariation scheme note that $\max_{\alpha \in \mathbb{A}_j} \tilde{\theta}^\alpha / \theta^\alpha \geq 1$ and $\min_{\alpha \in \mathbb{A}_j} \tilde{\theta}^\alpha / \theta^\alpha \leq 1$, for all $j \in [n]$. Thus

$$\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log \max_{\substack{j \in [n] \\ \alpha \in \mathbb{A}_j}} \frac{\tilde{\theta}^\alpha}{\theta^\alpha} - \log \min_{\substack{j \in [n] \\ \alpha \in \mathbb{A}_j}} \frac{\tilde{\theta}^\alpha}{\theta^\alpha}.$$

Now note that $\tilde{\theta} = \theta_{-j_s} \tilde{\theta}_{j_s}$, for a $\theta_{j_s} \in \Theta_j$, $j \in [n]$, since no two parameters in $\bigcup_{j \in [n]} \Theta_j$ can have exponent non-zero in the same monomial. Thus $\tilde{\theta}^\alpha / \theta^\alpha = \tilde{\theta}_{j_s} / \theta_{j_s}$ since $\alpha \in \{0, 1\}^k$ and the result follows. \square

We can now prove that the proportional covariation scheme is optimal for single full CPT analyses. This is important since a set of parameters might be varied to change an uncalibrated probability of interest, but a user might want to achieve this by choosing a distribution as close as possible to the original one. Several authors have posed this problem for BNs without finding a definitive answer [10, 45]. Here, exploiting our polynomial model representation, we can prove the optimality of the proportional scheme not only for BN models, but also for multilinear ones in single full CPT analyses.

Theorem 1. *Under the conditions of Proposition 4 and proportional covariations $\sigma_j(\theta_{j_s}, \theta_{j_i})$, the CD distance between \mathbf{p}_θ and $\mathbf{p}_{\tilde{\theta}}$ is minimized and can be written in closed form as*

$$\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log \max_{j \in [n]} \left\{ \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}}, \frac{1 - \tilde{\theta}_{j_i}}{1 - \theta_{j_i}} \right\} - \log \min_{j \in [n]} \left\{ \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}}, \frac{1 - \tilde{\theta}_{j_i}}{1 - \theta_{j_i}} \right\}. \quad (12)$$

Proof. First note that we can write equation (11) as

$$\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log \max \left\{ \max_{s \in [r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \dots, \max_{s \in [r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}} \right\} - \log \min \left\{ \min_{s \in [r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \dots, \min_{s \in [r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}} \right\}. \quad (13)$$

Now, let $\bar{\theta}_{j_i} = \tilde{\theta}_{j_i}$ and suppose $\bar{\theta}_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ is obtained via a valid covariation scheme, $j \in [n]$, $s \in [r_j]$. We want to prove that $\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) \geq \mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}})$. Suppose now the proportional scheme is optimal for one-way sensitivity analyses. If this is true, we must have that, for all $j \in [n]$,

$$\max_{s \in [r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} \geq \max_{s \in [r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}, \quad \text{and} \quad \min_{s \in [r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} \leq \min_{s \in [r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}.$$

Therefore,

$$\max \left\{ \max_{s \in [r_1]} \frac{\bar{\theta}_{1_s}}{\theta_{1_s}}, \dots, \max_{s \in [r_n]} \frac{\bar{\theta}_{n_s}}{\theta_{n_s}} \right\} \geq \max \left\{ \max_{s \in [r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \dots, \max_{s \in [r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}} \right\},$$

and

$$\min \left\{ \min_{s \in [r_1]} \frac{\bar{\theta}_{1_s}}{\theta_{1_s}}, \dots, \min_{s \in [r_n]} \frac{\bar{\theta}_{n_s}}{\theta_{n_s}} \right\} \leq \min \left\{ \min_{s \in [r_1]} \frac{\tilde{\theta}_{1_s}}{\theta_{1_s}}, \dots, \min_{s \in [r_n]} \frac{\tilde{\theta}_{n_s}}{\theta_{n_s}} \right\},$$

from which the optimality condition follows.

We thus have to prove that for a single parameter change, the proportional covariation scheme minimizes the CD distance in any multilinear model. The proof follows similar steps to the one in [9] for BNs. Fix $j \in [n]$ and note that if either $\theta_{j_i} = 0$ or $\theta_{j_i} = 1$ then the distance is infinite under both covariation schemes and the result holds. Consider now $\theta_{j_i} \in (0, 1)$ and suppose $\bar{\theta}_{j_i} = \tilde{\theta}_{j_i} > \theta_{j_i}$. Under a proportional scheme, we have that

$$\max_{s \in [r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} = \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}} \quad \text{and} \quad \min_{s \in [r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} = \frac{\theta_{j_s}(1 - \tilde{\theta}_{j_i})}{(\theta_{j_s}(1 - \theta_{j_i}))} = \frac{(1 - \tilde{\theta}_{j_i})}{(1 - \theta_{j_i})}.$$

Conversely, for the generic covariation scheme $\sigma(\theta_{j_s}, \bar{\theta}_{j_i})$ we have that

$$\begin{aligned} \frac{1 - \bar{\theta}_{j_i}}{1 - \theta_{j_i}} &= \frac{\sum_{s \in [r_j] \setminus \{j_i\}} \bar{\theta}_{j_s}}{\sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s}} = \frac{\sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s} (\bar{\theta}_{j_s} / \theta_{j_s})}{\sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s}} \\ &\geq \frac{\sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s} (\min_{k \in [r_j]} \bar{\theta}_k / \theta_k)}{\sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s}} = \min_{s \in [r_j]} \frac{\bar{\theta}_s}{\theta_s}. \end{aligned}$$

Thus since $(1 - \bar{\theta}_{j_i}) / (1 - \theta_{j_i}) = (1 - \tilde{\theta}_{j_i}) / (1 - \theta_{j_i})$ we have that $\min_{s \in [r_j]} \bar{\theta}_{j_s} / \theta_{j_s} \geq \min_{s \in [r_j]} \tilde{\theta}_{j_s} / \theta_{j_s}$. Furthermore,

$$\max_{s \in [r_j]} \frac{\bar{\theta}_{j_s}}{\theta_{j_s}} \geq \frac{\bar{\theta}_{j_i}}{\theta_{j_i}} = \frac{\tilde{\theta}_{j_i}}{\theta_{j_i}} = \max_{s \in [r_j]} \frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}.$$

It then follows that $\mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\bar{\theta}}) \geq \mathcal{D}_{\text{CD}}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}})$ when $\tilde{\theta}_{j_i} > \theta_{j_i}$ for one-way analyses. For the case $\tilde{\theta}_{j_i} < \theta_{j_i}$ the proof mirrors the one presented here. The explicit form of the

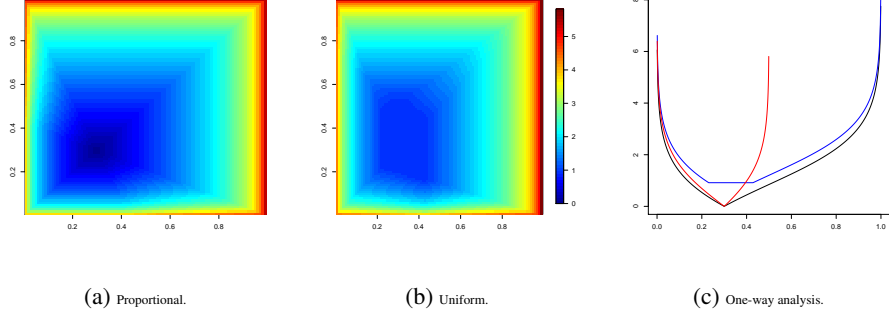


Figure 5: CD distances for Example 5 under different covariation schemes: proportional (black), uniform (blue), order-preserving (red).

distance under a proportional covariation schemes in equation (12) follows by noting that the maximum and the minimum can either be $\tilde{\theta}_{ji}/\theta_{ji}$ or $(1 - \tilde{\theta}_{ji})/(1 - \theta_{ji})$. \square

Example 5. In Figure 5 we plot the CD distance between the varied and the original probability distributions for Example 2 when θ_{21_1} (x-axis) and θ_{21_0} (y-axis in 5a and 5b) are varied for the covariation schemes so far considered. From Figures 5a and 5b we can get an intuition that the distance under proportional covariation is smaller than in the uniform case. This becomes clearer when we only let θ_{21_1} vary as shown in Figure 5c since the black line is always underneath the others.

4.4. ϕ -divergences

Although the CD distance is widely used in sensitivity analyses, comparisons between two generic distributions are usually performed by computing the KL divergence. For one-way sensitivity analysis in BNs, the KL divergence equals the KL divergence between the original and varied conditional probability distribution of the manipulated parameter times the marginal probability of the conditioning parent configuration. This means that one way sensitivity analyses based on KL distances can become computationally infeasible, since this constant term might need to be computed an arbitrary large number of times. In Proposition 5 below we demonstrate that this property is common to any ϕ -divergence for any multilinear model and single full CPT analyses.

Proposition 5. Let $\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}} \in \mathbb{P}_\Psi$, where \mathbb{P}_Ψ is a multilinear parametric model and $\mathbf{p}_{\tilde{\theta}}$ arises from \mathbf{p}_θ by varying θ_{j_i} to $\tilde{\theta}_{j_i}$ and $\theta_{j_s} \in \Theta_j \setminus \{\theta_{j_i}\}$ to $\tilde{\theta}_{j_s} = \sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$, where $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i})$ is a valid covariation scheme, $j \in [n], s \in [r_j] \setminus \{j_i\}$. Then the ϕ -divergence between $\mathbf{p}_{\tilde{\theta}}$ and \mathbf{p}_θ is equal to

$$\mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}, \mathbf{p}_\theta) = \sum_{j \in [n]} \mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}^j, \mathbf{p}_\theta^j) \sum_{s \in [r_j]} \sum_{\alpha \in \mathbb{A}_{-j_s}} \theta_{-j_s}^\alpha, \quad (14)$$

where \mathbf{p}_θ^j denotes the vector of atomic probabilities in Θ_j .

Proof. For a model with monomial parametrisation the ϕ -divergence can be written as

$$\mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}, \mathbf{p}_\theta) = \sum_{\alpha \in \mathbb{A}} \theta^\alpha \phi\left(\frac{\tilde{\theta}^\alpha}{\theta^\alpha}\right) = \sum_{j \in [n]} \sum_{\alpha \in \mathbb{A}_j} \theta^\alpha \phi\left(\frac{\tilde{\theta}^\alpha}{\theta^\alpha}\right) + \sum_{\alpha \in \mathbb{A} \setminus \bigcup_{j \in [n]} \mathbb{A}_j} \theta^\alpha \phi\left(\frac{\tilde{\theta}^\alpha}{\theta^\alpha}\right).$$

Notice that for $\alpha \in \mathbb{A} \setminus \bigcup_{j \in [n]} \mathbb{A}_j$, $\tilde{\theta}^\alpha / \theta^\alpha = 1$. Thus, since $\phi(1) = 0$, we then have that

$$\begin{aligned} \mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}, \mathbf{p}_\theta) &= \sum_{j \in [n]} \sum_{\alpha \in \mathbb{A}_j} \theta^\alpha \phi\left(\frac{\tilde{\theta}^\alpha}{\theta^\alpha}\right) = \sum_{j \in [n]} \sum_{s \in [r_j]} \sum_{\alpha \in \mathbb{A}_{-j_s}} \theta_{-j_s}^\alpha \theta_{j_s} \phi\left(\frac{\theta_{-j_s}^\alpha \tilde{\theta}_{j_s}}{\theta_{-j_s}^\alpha \theta_{j_s}}\right) \\ &= \sum_{j \in [n]} \sum_{s \in [r_j]} \theta_{j_s} \phi\left(\frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}\right) \sum_{\alpha \in \mathbb{A}_{-j_s}} \theta_{-j_s}^\alpha = \sum_{j \in [n]} \mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}^j, \mathbf{p}_\theta^j) \sum_{s \in [r_j]} \sum_{\alpha \in \mathbb{A}_{-j_s}} \theta_{-j_s}^\alpha. \end{aligned}$$

□

The additional complexity of having to compute the constant term in equation (14) has limited the use of KL divergences, and more generally ϕ -divergences, in both practical and theoretical sensitivity investigations in discrete BNs. However, looking at probabilistic models from a polynomial point of view, we are able here to establish an additional strong theoretical justification for the use proportional covariation even in single full CPT analyses, since this also minimizes any ϕ -divergence.

Theorem 2. Under the conditions of Proposition 5, $\mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}, \mathbf{p}_\theta)$ is minimized by the proportional covariation schemes $\sigma_j(\theta_{j_s}, \tilde{\theta}_{j_i}) = (1 - \tilde{\theta}_{j_i})\theta_{j_s} / (1 - \theta_{j_i})$.

Proof. Since $\sum_{s \in [r_j]} \sum_{\alpha \in \mathbb{A}_{-j_s}} \theta_{-j_s}^\alpha$ in equation (14) is a positive constant, $\mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}, \mathbf{p}_\theta)$ is minimized if each $\mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}^j, \mathbf{p}_\theta^j)$ attains its minimum. Fix a $j \in [n]$. We use the method of Lagrange multipliers to demonstrate that $\mathcal{D}_\phi(\mathbf{p}_{\tilde{\theta}}^j, \mathbf{p}_\theta^j)$ is minimized by proportional

covariation, subject to the constraint that $\sum_{s \in [r_j]} \tilde{\theta}_{j_s} - 1 = 0$. Define

$$L = \sum_{s \in [r_j]} \theta_{j_s} \phi\left(\frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}\right) - \lambda \left(\sum_{s \in [r_j]} \tilde{\theta}_{j_s} - 1 \right).$$

Taking the first derivative of L with respect to $\tilde{\theta}_{j_s}$ and equating it to zero gives

$$\frac{\partial}{\partial \tilde{\theta}_{j_s}} L = \phi'\left(\frac{\tilde{\theta}_{j_s}}{\theta_{j_s}}\right) = \lambda,$$

where ϕ' denotes the derivative of ϕ . By inverting we then deduce that

$$\tilde{\theta}_{j_s} = \phi'(\lambda)^{-1} \theta_{j_s}. \quad (15)$$

Since equation (15) holds for every $s \in [r_j] \setminus \{j_i\}$ we have that

$$\sum_{s \in [r_j] \setminus \{j_i\}} \tilde{\theta}_{j_s} = \phi'(\lambda)^{-1} \sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s} \quad (16)$$

Now take the first partial derivative of L with respect to λ and equate it to zero. This gives

$$\frac{\partial}{\partial \lambda} L = \sum_{s \in [r_j]} \tilde{\theta}_{j_s} = 1 \implies \sum_{s \in [r_j] \setminus \{j_i\}} \tilde{\theta}_{j_s} = 1 - \tilde{\theta}_{j_i} \quad (17)$$

Plugging the right hand side of (17) into (16), we deduce that

$$\phi'(\lambda)^{-1} = \frac{1 - \tilde{\theta}_{j_i}}{\sum_{s \in [r_j] \setminus \{j_i\}} \theta_{j_s}} = \frac{1 - \tilde{\theta}_{j_i}}{1 - \theta_{j_i}}. \quad (18)$$

Thus, by plugging (18) into (15) we conclude that

$$\tilde{\theta}_{j_s} = \frac{1 - \tilde{\theta}_{j_i}}{1 - \theta_{j_i}} \theta_{j_s}.$$

This is guaranteed to be a minimum by the convexity of the function ϕ . \square

Example 6. As an example of a ϕ -divergence, in Figure 6 we plot $\text{KL}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}})$ for Example 2 when $\theta_{2_1 1_1}$ (x-axis) and $\theta_{2_1 0}$ (y-axis in 6a and 6b) are varied for the covariation schemes so far considered. From Figures 6a and 6b we can get an intuition that the KL divergence under proportional covariation is smaller than in the uniform case. This becomes clearer when we only let $\theta_{2_1 1_1}$ vary as shown in Figure 6c since the black line is always underneath the others.

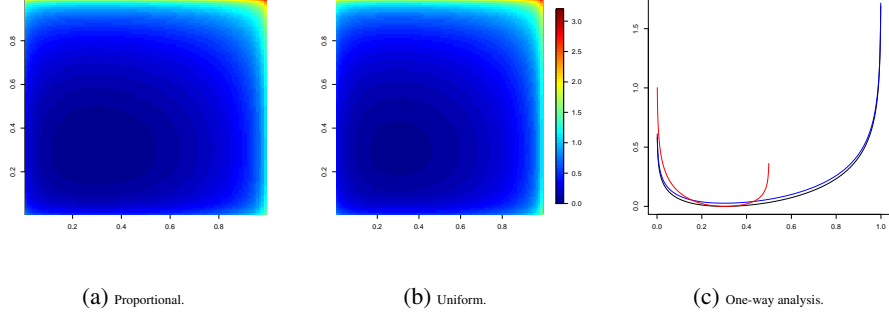


Figure 6: KL divergences for Example 6 under different covariation schemes: proportional (black), uniform (blue), order-preserving (red).

5. One-way sensitivity analysis in non-multilinear models

For multilinear models we have been able to provide a unifying framework to perform various sensitivity analyses by deducing closed forms for both sensitivity functions and various divergences. Unfortunately, this is not possible for non-multilinear parametric models because these will depend on the degree of both the indeterminate to be varied and the covaried parameters, which is not necessarily equal to one. However, representing the model through its interpolating polynomial will enable us to study central properties of various sensitivity analyses. Since sensitivity functions and CD distances are most commonly applied, in this section we will only focus on these. Furthermore, because of the much more general structure underlying non-multilinear models, we will restrict our discussion to one-way sensitivity methods. As in Section 4.1, we let $\theta_i \in \Theta_C$ be varied to $\tilde{\theta}_i$, where $\Theta_C = \{\theta_j : j \in [r]\}$ is the set of parameters including θ_i which need to respect the sum-to-one condition.

5.1. Sensitivity functions

As in Section 4 we let f_{y_T} denote a sensitivity function and $f_{y_T}^{y_o}$ a posterior sensitivity function.

Proposition 6. *Consider a parametric model \mathbb{P}_Ψ with monomial parametrisation Ψ . Let θ_i vary to $\tilde{\theta}_i$ and $\theta_j \in \Theta_C \setminus \{\theta_i\}$ covary according to a linear scheme. The sensitivity function $f_{y_T}(\tilde{\theta}_i)$ is then a polynomial with degree lower or equal to d , where*

$\hat{\theta}_{1_1 2_0} = 0.4,$	$\hat{\theta}_{1_1 2_1} = 0.6,$	$\hat{\theta}_{1_1 2_2} = 0.7,$	$\hat{\theta}_{3_0 2_0} = 0.9,$	$\hat{\theta}_{3_0 2_1} = 0.6,$	$\hat{\theta}_{3_0 2_2} = 0.2,$
$\hat{\theta}_{2_1 2_0 3_0} = 0.3,$	$\hat{\theta}_{2_2 2_0 3_0} = 0.2$	$\hat{\theta}_{2_1 2_1 3_0} = 0.3$	$\hat{\theta}_{2_2 2_1 3_0} = 0.5$	$\hat{\theta}_{2_1 2_2 3_0} = 0.5$	$\hat{\theta}_{2_2 2_2 3_0} = 0.3.$

Table 3: Probability specifications for Example 7.

$d = \max_{\alpha \in \mathbb{T}_C, j \in [r]} \{\alpha_j\}$. The posterior sensitivity function $f_{y_T}^{y_0}(\tilde{\theta}_i)$ is a rational function whose numerator and denominator are polynomials again with degree lower or equal to d .

Proof. This follows by noting that under the conditions of the proposition an exponent $\alpha_{i,y}$ in Definition 3 of monomial parametrisation, for any $\alpha_y \in \mathbb{A}$ cannot be larger than d . \square

Similar results to Proposition 6 were presented in [13, 14, 44] for some specific classes of DBNs only.

The specific form of the sensitivity function in non-multilinear models depends both on the form of the interpolating polynomial, on the parameter to be varied and on the parameters that are consequently covaried. Therefore it is not possible to deduce a unique closed form expression since this will explicitly depend on the degree of all the above indeterminates. However, the interpolating polynomial enables us to identify a straightforward procedure to compute $f_{y_T}(\tilde{\theta}_i)$ as follows:

1. determine the polynomial $c_{\mathbb{P}_\Psi}(\theta, y_T)$ for an event \mathbb{Y}_T ;
2. replace θ_i by $\tilde{\theta}_i$;
3. replace θ_j by $\sigma(\theta_j, \tilde{\theta}_i)$ for $\theta_j \in \Theta_C \setminus \{\theta_i\}$.

Example 7. Suppose the definition of the DBN model in Example 3 is embellished by the probability specifications in Table 3. Suppose further the first time point distribution coincides with one defined in Table 1 for the context-specific BN in Example 2. We are now interested in the event $(Y_1(t) = 1, Y_2(1) = 1, Y_2(4) = 1, Y_3(t) = 0, t \in [4])$. By applying the above procedure to the interpolating polynomial of this DBN we can deduce that the sensitivity function when $\hat{\theta}_{2_1 2_1 3_0}$ is varied to x equals

$$ax^3 + bx\sigma(\hat{\theta}_{2_0 2_1 3_0}, x) + cx\sigma(\hat{\theta}_{2_2 2_1 3_0}, x) + d\sigma(\hat{\theta}_{2_0 2_1 3_0}, x) + e\sigma(\hat{\theta}_{2_2 2_1 3_0}, x),$$

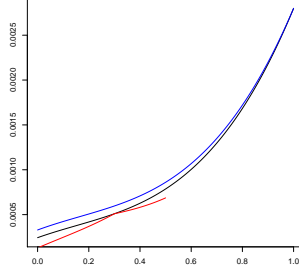


Figure 7: Sensitivity functions in Example 7 when $\hat{\theta}_{2,2,3_0}$ (on the x-axis) is varied: proportional (black), uniform (blue) and order-preserving (red).

where $a, b, c, d, e \in [0, 1]$. This is plotted in Figure 7 for different covariation schemes. As formalized in Proposition 6 these are not linear in their arguments, but more generally polynomial. As in the multilinear case, we can notice that the probability of interest under order-preserving covariation behaves rather differently than under uniform and proportional covariation.

5.2. CD distance

We next introduce a general procedure to compute the CD distance for parametric models with monomial parametrisation.

Proposition 7. *Let $\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}} \in \mathbb{P}_\Psi$, where \mathbb{P}_Ψ is a parametric model with monomial parametrisation. The CD distance between \mathbf{p}_θ and $\mathbf{p}_{\tilde{\theta}}$ when θ_i is varied to $\tilde{\theta}_i$ and $\theta_j \in \Theta_C \setminus \{\theta_i\}$ is covaried to $\tilde{\theta}_j = \sigma(\theta_j, \tilde{\theta}_i)$, $j \in [r] \setminus \{i\}$, according to a valid covariation scheme can be computed from the interpolating polynomial as follows:*

1. set $\theta^\alpha = 0$ for all $\alpha \in \mathbb{A} \setminus \mathbb{A}_C$;
2. set $\theta_k = 1$, if $k \notin [r]$, in any monomial θ^α , $\alpha \in \mathbb{A}_C$;
3. call Φ the set of remaining monomials, ϕ^α a generic element of Φ and $\tilde{\phi}^\alpha$ its varied version;
4. set $u = \max_\Phi \tilde{\phi}^\alpha / \phi^\alpha$ and $l = \min_\Phi \tilde{\phi}^\alpha / \phi^\alpha$;
5. compute $\text{CD}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log(u) - \log(l)$.

It then follows that $\text{CD}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \text{CD}(\mathbf{p}_\phi, \mathbf{p}_{\tilde{\phi}})$, where \mathbf{p}_ϕ and $\mathbf{p}_{\tilde{\phi}}$ denote respectively the original and the varied probability mass function of the parametric model whose atomic probabilities are the elements of Φ .

Proof. In a multilinear parametric model each atom is associated to a monomial θ^α . For all $\alpha \in \mathbb{A} \setminus \mathbb{A}_C$, we have that

$$\frac{p_{\tilde{\theta}}(\mathbf{y})}{p_\theta(\mathbf{y})} = \frac{\tilde{\theta}^\alpha}{\theta^\alpha} = 1,$$

and there will always be ratios smaller or bigger than one because of the validity of the covariation scheme. Therefore, these monomials have no impact on the distance (step 1). For $\alpha \in \mathbb{A}_C$, we have that

$$\frac{\tilde{\theta}^\alpha}{\theta^\alpha} = \frac{\tilde{\theta}_C^{\alpha_C}}{\theta_C^{\alpha_C}} \triangleq \frac{\tilde{\phi}^\alpha}{\phi^\alpha},$$

where $\theta_C = \prod_{j \in [r]} \theta_j$ and $\alpha_C \in \mathbb{N}_0^r$ is the associated exponent vector. Therefore the distance depends only on the monomials computed in steps 2 and 3. The result then follows from the definition of the CD distance. \square

Example 8. By following the procedure in Proposition 7 we deduce that the set Φ for the DBN model in Example 3 when $\hat{\theta}_{2_1 2_1 3_0}$ is varied equals

$$\Phi = \left\{ \hat{\theta}_{2_1 2_1 3_0}^3, \hat{\theta}_{2_1 2_1 3_0}^2 \hat{\theta}_{2_0 2_1 3_0}, \hat{\theta}_{2_1 2_1 3_0}^2 \hat{\theta}_{2_2 2_1 3_0} \hat{\theta}_{2_1 2_1 3_0}^2, \hat{\theta}_{2_0 2_1 3_0}^2, \hat{\theta}_{2_2 2_1 3_0}^2, \right. \\ \left. \hat{\theta}_{2_1 2_1 3_0} \hat{\theta}_{2_0 2_1 3_0}, \hat{\theta}_{2_1 2_1 3_0} \hat{\theta}_{2_2 2_1 3_0}, \hat{\theta}_{2_0 2_1 3_0} \hat{\theta}_{2_2 2_1 3_0}, \hat{\theta}_{2_1 2_1 3_0}, \hat{\theta}_{2_0 2_1 3_0}, \hat{\theta}_{2_2 2_1 3_0} \right\}$$

The algorithm selects the maximum ratio and the minimum ratio between any of these monomials and their non-varied versions, and then use these in the standard formula of the CD distance. In Figure 8 we plot the CD distance as a function of $\hat{\theta}_{2_1 2_1 3_0}$ under different covariation schemes. This shows that the distance is smallest for the proportional covariation scheme. In Theorem 1 we showed that this is the case for single full CPT sensitivity analyses in multilinear models. Unfortunately, this result does not hold in the non-multilinear case as shown in the following example.

Example 9. Consider two random variables Y_1 and Y_2 and suppose $\mathbb{Y}_1 = \mathbb{Y}_2 = [3]$. Suppose also

$$\theta_{1i} = \Pr(Y_1 = i) = \Pr(Y_2 = i | Y_1 = j), \quad i \in [3], j \in [2].$$

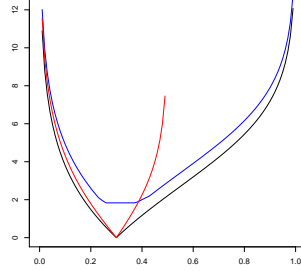


Figure 8: CD distance in Example 8 when $\hat{\theta}_{2,2,3_0}$ (on the x-axis) is varied: proportional (black), uniform (blue) and order-preserving (red).

Suppose we let θ_{11} vary and θ_{12} and θ_{13} covary according to a valid scheme. The set Φ of Proposition 7 is then equal $\{\theta_{11}^2, \theta_{11}\theta_{12}, \theta_{13}, \theta_{12}^2, \theta_{11}\theta_{13}, \theta_{12}\theta_{13}\}$. Suppose $\theta_{11} = 0.33$, $\theta_{12} = 0.33$, $\theta_{13} = 0.34$ and let θ_{11} be varied to 0.4. In this situation the CD distance under a proportional scheme is 2.52, whilst under a uniform scheme the distance equals 2.50. For this parameter variation, the uniform scheme would then be preferred to a proportional one if a user wishes to minimize the CD distance. Conversely, if θ_{11} is set to 0.2 the distance is smaller under the proportional scheme (2.89) than under the uniform one (2.92).

Therefore, whilst for multilinear models the choice of updating probabilities with a proportional covariation scheme can be justified in terms of some ‘optimality criterion’ based on the minimization of the CD distance, for non-multilinear models the choice of the covariation scheme becomes critical. Our examples demonstrated that inference can be greatly affected by the chosen covariation scheme. With no theoretical justification to use one over another, any output from such a sensitivity analysis of a non-multilinear model will be ambiguous unless a convincing rationale for the choice of the covariation scheme can be found.

We next deduce a closed form expression for the CD distance under the proportional covariation scheme.

Proposition 8. *Under the condition of Proposition 7 and assuming a proportional*

covariation scheme the CD distance is

$$\text{CD}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log \max_{\alpha \in \mathbb{A}_{-C}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|\alpha_{-i}|} \frac{\tilde{\theta}_i^{\alpha_i}}{\theta_i^{\alpha_i}} - \log \min_{\alpha \in \mathbb{A}_{-C}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|\alpha_{-i}|} \frac{\tilde{\theta}_i^{\alpha_i}}{\theta_i^{\alpha_i}}, \quad (19)$$

where $\mathbb{A}_{-C} \in \mathbb{N}_0^r$ is the set including the exponents in \mathbb{A}_C where the entries relative to indeterminates not in Θ_C are deleted, $\alpha_{-i} \in \mathbb{N}_0^{r-1}$ is an exponent where the entry relative to θ_i is deleted and $|\alpha_{-i}|$ is the sum of its entries.

Proof. From Proposition 7 we can write the CD distance as

$$\text{CD}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \log \max_{\alpha \in \mathbb{A}_{-C}} \frac{\tilde{\phi}^\alpha}{\phi^\alpha} - \log \min_{\alpha \in \mathbb{A}_{-C}} \frac{\tilde{\phi}^\alpha}{\phi^\alpha}.$$

Recall that under a proportional scheme an indeterminate $\theta_j \in \Theta_C \setminus \{\theta_i\}$ is varied to $\tilde{\theta}_j = \theta_j(1 - \tilde{\theta}_i)/(1 - \theta_i)$. We then have that, for any $\phi \in \Phi$,

$$\tilde{\phi}^\alpha = \tilde{\theta}_i^{\alpha_i} \prod_{j \in [r] \setminus \{i\}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \theta_j \right)^{\alpha_j} = \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|\alpha_{-i}|} \theta_{-i}^{\alpha_{-i}} \tilde{\theta}_i^{\alpha_i},$$

and therefore

$$\frac{\tilde{\phi}^\alpha}{\phi^\alpha} = \frac{\tilde{\theta}_i^{\alpha_i}}{\theta_i^{\alpha_i}} \left(\frac{1 - \tilde{\theta}_i}{1 - \theta_i} \right)^{|\alpha_{-i}|}.$$

Then, since $\text{CD}(\mathbf{p}_\theta, \mathbf{p}_{\tilde{\theta}}) = \text{CD}(\mathbf{p}_\phi, \mathbf{p}_{\tilde{\phi}})$ by Proposition 7, we deduce the closed form in equation (19) by substituting the above expression into the definition of CD distance. \square

Example 10. Consider again the monomial set Φ of Example 8 and assume the parameters $\hat{\theta}_{2_0 2_1 3_0}$ and $\hat{\theta}_{2_2 2_1 3_0}$ are covaried according to a proportional scheme. Call $\hat{\theta} = \hat{\theta}_{2_1 2_1 3_0}$ and let x be its varied version. From Proposition 8 we can deduce that the CD distance will then depend on the maximum and minimum value in the set of ratios

$$\left\{ \frac{x^3}{\hat{\theta}^3}, \frac{x^2}{\hat{\theta}^2}, \frac{x}{\hat{\theta}}, \frac{x^2}{\hat{\theta}^2} \left(\frac{1-x}{1-\hat{\theta}} \right), \frac{x}{\hat{\theta}} \left(\frac{1-x}{1-\hat{\theta}} \right), \left(\frac{1-x}{1-\hat{\theta}} \right)^2, \frac{1-x}{1-\hat{\theta}} \right\}.$$

6. Discussion

The definition of a parametric model by its interpolating polynomial has proven useful to investigate how changes in the input probabilities affect an output of interest. We have been able to demonstrate not only that standard results for one-way analyses

in BN models are valid for a large number of other models and single full CPT investigations, but also new theoretical justifications for the use of proportional covariation based on a variety of divergence measures. Then, the flexibility of the interpolating polynomial representation enabled us to investigate an even larger class of models, for instance DBNs, extending sensitivity methods to dynamic settings. In this framework both sensitivity functions and CD distances exhibit different properties than in the simpler multilinear case, with the potential of even more informative sensitivity investigations. Importantly, we have been able to produce a new fast procedure to compute the CD distance in non-multilinear models.

Having demonstrated the usefulness of our polynomial approach in single full CPT analyses, we next plan to address the rather more complicated situation of generic multi-way analyses. In particular by representing probabilities in terms of monomials we can relate multi-way analyses in multilinear models to one-way sensitivities in non-multilinear ones. It can be seen that sensitivity functions for multi-way analyses will not simply be multilinear but also include interaction terms. Similarly, the CD distance will be affected by such interactions and not simply correspond to the CD distance of the appropriate CPT. Example 9 would therefore suggest that the proportional covariation scheme is not optimal in this context. However, we notice that the monomials from multi-way analyses in multilinear models are a subset of those arising from non-multilinear ones. Although these can be of an arbitrary degree, each indeterminate of the monomial will have exponent one by construction. Therefore, there is no conclusive proof of the non-optimality of the proportional scheme in generic multi-way analyses for multilinear models and BNs. However the polynomial representation of probabilities in BNs and related models gives us a promising starting point to start investigating this class of problems.

Acknowledgements

M. Leonelli was supported by Capes, C. Görgen was supported by the EPSRC grant EP/L505110/1 and J. Q. Smith was supported by the EPSRC grant EP/K039628/1.

References

References

- [1] P. A. Aguilera, A. Fernández, R. Fernández, R. Rumí, and A. Salmerón. Bayesian networks in environmental modelling. *Environ. Modell. Softw.*, 26:1376–1388, 2011.
- [2] S. M. Ali and S. D. Silvey. A general class of coefficients of divergence of one distribution from another. *J. Roy. Statist. Soc. Ser. B*, 28:131–142, 1966.
- [3] O. Anacleto and C. M. Queen. Dynamic chain graph models for multivariate time series. Technical report, Department of Statistics, Open University, 2013.
- [4] L. M. Barclay, R. A. Collazo, J. Q. Smith, P. A. Thwaites, and A. E. Nicholson. The dynamic chain event graph. *Electron. J. Stat.*, 9:2130–2169, 2015.
- [5] J. H. Bolt and L. C. van der Gaag. Balanced tuning of multi-dimensional Bayesian network classifiers. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 210–220. Springer, 2015.
- [6] C. Boutilier, N. Friedman, M. Goldszmidt, and D. Koller. Context-specific independence in Bayesian networks. In *Proceedings of the 12th Conference on Uncertainty in Artificial Intelligence*, pages 115–123, 1996.
- [7] R. Cano, C. Sordo, and J. M. Gutiérrez. Applications of Bayesian networks in meteorology. In *Advances in Bayesian Networks*, pages 309–328. Springer, 2004.
- [8] E. Castillo, J. M. Gutiérrez, and A. S. Hadi. Sensitivity analysis in discrete Bayesian networks. *IEEE T. Syst. Man Cyb.*, 27:412–423, 1997.
- [9] H. Chan and A. Darwiche. When do numbers really matter? *J. Artificial Intelligence Res.*, 17:265–287, 2002.
- [10] H. Chan and A. Darwiche. Sensitivity analysis in Bayesian networks: from single to multiple parameters. In *Proceedings of the 16th Conference on Uncertainty in Artificial Intelligence*, pages 317–325, 2004.

- [11] H. Chan and A. Darwiche. A distance measure for bounding probabilistic belief change. *Internat. J. Approx. Reason.*, 38:149–174, 2005.
- [12] H. Chan and A. Darwiche. Sensitivity analysis in Markov networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, pages 1300–1305, 2005.
- [13] T. Charitos and L. C. van der Gaag. Sensitivity analysis of Markovian models. In *Proceedings of the FLAIRS Conference*, pages 806–811, 2006.
- [14] T. Charitos and L. C. van der Gaag. Sensitivity analysis for threshold decision making with dynamic networks. In *Proceedings of the 22nd Conference on Uncertainty in Artificial Intelligence*, pages 72–79, 2006.
- [15] S. H. Chen and C. A. Pollino. Good practice in Bayesian network modelling. *Environ. Modell. Softw.*, 37:134–145, 2012.
- [16] V. M. H. Coupé and L. C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Ann. Math. Artif. Intell.*, 36:323–356, 2002.
- [17] R. G. Cowell, A. P. Dawid, Lauritzen S. L., and D. J. Spiegelhalter. *Probabilistic networks and expert systems*. Springer-Verlag, New York, 1999.
- [18] I. Csiszár. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizität von markoffschen ketten. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 8:85–108, 1963.
- [19] P. Dagum and E. Horvitz. A Bayesian analysis of simulation algorithms for inference in belief networks. *Networks*, 23:499–516, 1993.
- [20] A. Darwiche. A differential approach to inference in Bayesian networks. *J. ACM*, 3:280–305, 2003.
- [21] M. Drton, B. Sturmfels, and S. Sullivant. *Lectures on algebraic statistics*. Birkhäuser Verlag, Basel, 2009.

- [22] S. French. Modelling, making inferences and making decisions: the roles of sensitivity analysis. *Top*, 11:229–251, 2003.
- [23] M. A. Gómez-Villegas, P. Main, and R. Susi. Sensitivity analysis in Gaussian Bayesian networks using a divergence measure. *Comm. Statist. Theory Methods*, 36:523–539, 2007.
- [24] M. A. Gómez-Villegas, P. Main, and R. Susi. The effect of block parameter perturbations in Gaussian Bayesian networks: sensitivity and robustness. *Inform. Sci.*, 222:429–458, 2013.
- [25] C. Görgen and J. Q. Smith. Equivalence classes of staged trees. Technical report, CRISM 15-12, University of Warwick, 2015.
- [26] C. Görgen, M. Leonelli, and J. Q. Smith. A differential approach for staged trees. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 346–355. Springer, 2015.
- [27] D. Heckerman, A. Mamdani, and M. P. Wellman. Real-world applications of Bayesian networks. *Commun. ACM*, 38:24–26, 1995.
- [28] H. Jeffreys. An invariant form for the prior probability in estimation problems. *Proc. Roy. Soc. London Ser. A*, 186:453–461, 1946.
- [29] M. I. Jordan. Graphical models. *Statist. Sci.*, 19:140–155, 2004.
- [30] D. Koller and U. Lerner. Sampling in factored dynamic systems. In *Sequential Monte Carlo Methods in Practice*, pages 445–464. Springer, 2001.
- [31] K. Korb and A. E. Nicholson. *Bayesian artificial intelligence*. CRC Press, Boca Raton, 2010.
- [32] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statistics*, 22:79–86, 1951.
- [33] S. L. Lauritzen. *Graphical models*. Oxford University Press, Oxford, 1996.

- [34] S. Low-Choy, A. James, J. Murray, and K. Mengersen. Elicitor: a user-friendly interactive tool to support scenario-based elicitation of expert knowledge. In *Expert Knowledge and its Application in Landscape Ecology*, pages 39–67. Springer, 2012.
- [35] K. P. Murphy. *Dynamic Bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.
- [36] M. Neil, N. Fenton, and L. Nielson. Building large-scale Bayesian networks. *Knowl. Eng. Rev.*, 15:257–284, 2000.
- [37] L. Pardo. *Statistical inference based on divergence measures*. CRC Press, Boca Raton, 2005.
- [38] J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan-Kaufman, San Francisco, 1988.
- [39] J. Pensar, H. Nyman, J. Lintusaari, and J. Corander. The role of local partial independence in learning of Bayesian networks. *Intern. J. Approx. Reason.*, 69: 91–105, 2016.
- [40] G. Pistone, E. Riccomagno, and H. P. Wyn. Gröbner bases and factorisation in discrete probability and Bayes. *Stat. Comput.*, 11:37–46, 2001.
- [41] J. Pitchforth and K. Mengersen. A proposed validation framework for expert elicited Bayesian networks. *Expert Syst. Appl.*, 40:162–167, 2013.
- [42] C. A. Pollino, O. Woodberry, A. Nicholson, K. Korb, and B. T. Hart. Parameterisation and evaluation of a Bayesian network for use in an ecological risk assessment. *Environ. Modell. Softw.*, 22:1140–1152, 2007.
- [43] E. Rajabally, P. Sen, S. Whittle, and J. Dalton. Aids to Bayesian belief network construction. In *Proceedings of the 2nd International Conference on Intelligence Systems*, pages 457–461, 2004.
- [44] S. Renooij. Efficient sensitivity analysis in hidden Markov models. *Internat. J. Approx. Reason.*, 53:1397–1414, 2012.

- [45] S. Renooij. Co-variation for sensitivity analysis in bayesian networks: properties, consequences and alternatives. *Internat. J. Approx. Reason.*, 55:1022–1042, 2014.
- [46] E. Riccomagno. A short history of algebraic statistics. *Metrika*, 69:397–418, 2009.
- [47] J. Q. Smith and P. E. Anderson. Conditional independence and chain event graphs. *Artificial Intelligence*, 172:42–68, 2008.
- [48] L. C. van der Gaag, S. Renooij, and V. M. H. Coupé. Sensitivity analysis of probabilistic networks. In *Advances in Probabilistic Graphical Models*, pages 103–124. Springer, 2007.